International Journal of Advances in Intelligent Informatics
Vol. 3, No. 2, July 2017, pp. 98-106
ISSN: 2442-6571

Spatial Data Modeling in Disposable Income Per Capita in China using Nationwide Spatial Autoregressive (SAR)

Tuti Purwaningsih [a,1,*], Anusua Ghosh [b,1], Chumairoh Chumairoh [a,2]

[a] *Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia*
[b] *University of South Australia, Australia*
[1] *tuti.purwaningsih@uii.ac.id *;* [2] *anusua.ghosh@mymail.unisa.edu.au;* [3] *chumairohazzahra@yahoo.com*
** corresponding author*

ARTICLE INFO

Article history:
Received July 22, 2017
Revised August 21, 2017
Accepted August 21, 2017

Keywords:
Per Capita Disposable Income Nationwide
Industry
Tourism
Spatial Autoregressive Model (SAR)

ABSTRACT

China as a country became the economic center of the world. However, with a population of 1.3 billion, China's per capita income is still at number 80 in the world. In the world, considering the imbalance between town and country with 100 million people still living in poverty. Thus, to address this imbalance, it is necessary to study the condition in depth, because income per capita is often used as a benchmark to measure the prosperity of a country. With greater and equitable income per capita, the country will be judged increasingly affluent. Two factors, mainly industry and tourism, play an important role in the economic progress in China. These are include Per capita Disposable Income Nationwide (yuan), Total Value of Exports of operating units (1,000 USD), Registered Unemployed Person in Urban Area (10000 person), Foreign Exchange Earning from International tourism(in millions USD) and Number of Overseas Visitor Arrivals (million person/time). Thus, it is necessary to investigate the influence of these factors to increase per capita income. Since the economic development of a region usually affect the surrounding area, this study aims to include spatial effects, using Spatial Autoregressive (SAR) Model. The results suggest that the per capita income affected by the Tourism factor is about 58.65% (R-squared).

I. Introduction

China is a country with an advanced economy, includes in the list of countries which has the biggest export, and China's economic strength is predicted that will defeat the United States [1]. In 2014, the regional director for GBTA (Global Business Travel Association) reported that China's economic growth also encouraged the tourist's business sector. Moreover, in the same year, China had made progress (economic) super fast with the value of GDP for 2014 was 28.3-fold rise and per capita rise 19-fold. Revitalization of the Chinese nation to make China's large emerging economies at the center of both worlds. However, with a population of 1.3 billion, China's per capita income is still at number 80 in the world, where 100 million people are still poor and are not in balance between town and country. With the advancement of a famous industry rapid development, China's per capita income was still not balanced between urban and rural areas; It is necessary assessments rely more deeply to solve it because after income per capita is often used as a benchmark for the prosperity of a country. If income per capita is greatest, the country will be judged increasingly affluent. Moreover, China belongs to the part of developed countries [1].

The understanding of factors that influence the per capita income is certainly very important so that It might be used as reference in decision-making in determining which factors greatly contribute to greater per capita income. It can be used as a strong basis in determining a policy that will be taken and is expected to facilitate the making of a policy so that the various possibilities that may occur regarding loss or weakness can be overcome. Also, tourism industry has a close relationship in advancing the economy in China. Therefore, the researchers aimed to examine how the influence of

industry and tourism on the economy, particularly its effect on per capita income. The form of analysis is using spatial regression analysis. It is based on the theory advanced by Tobbler that everything is interconnected with each other, but something closer more influence [2]. Their spatial effects are a common problem among the regions and especially the regions adjacent to each other [3]. With this is expected to be found the significant factors that affect the income per capita the research could found the measurement to improve the welfare of the community equally.

II. Literatur Review

A. Spatial Statistics

Spatial statistics is a statistical method used to analyze spatial data. Spatial data is data that contains information "location," so not only "what" measurable but indicates the location where the data is located. Spatial data may include information regarding the geographic location such as the location of the latitude and longitude of each border region and between regions. Simply put spatial data expressed as the address information. In another form, spatial data is expressed in the form of grid coordinates as in the grain map or the form of pixels as in the form of satellite imagery. Thus the approach of spatial statistical analysis is usually presented in the form of thematic maps [2].

B. Spatial Data Analysis

Spatial data is data that contains the location or geographical information of a region. Spatial analysis leads to many operations and concepts including simple calculations, classifications, structuring, geometric overlap, and cartographic modeling [4]–[6]. In general, spatial analysis requires data based on the location and contains the characteristics of the location. Spatial analysis consists of three groups namely visualization, exploration, and modeling. Visualization is to inform the results of spatial analysis. Exploration is to process spatial data with statistical methods. While modeling is showing the existence of the concept of causality by using methods from spatial data sources and nonspatial data to predict the existence of spatial patterns [7], [8]. Locations in spatial data should be measured to be aware of any spatial effects that occur. Location information can be identified from two sources [9]:

1. Neighborhood relations. The neighboring relationship reflects the relative location of one spatial unit or location to another in a given space. The neighboring relationships of the spatial units are usually formed on the map. The neighborhood of these spatial units is expected to reflect a high degree of spatial dependence when compared to spatially located units that are located far apart.
2. Distance (distance). Location in a certain space with the latitude and longitude into a source of information. This information is used to calculate the distance between the points contained in space. It is expected that the strength of spatial dependence will decrease according to the distance.

C. Spatial Autocorrelation

Spatial autocorrelation is an estimate of the correlation between the value of observations relating to spatial locations at the same variable. Positive spatial autocorrelation shows the similarity value from adjacent locations and tend to cluster. Negative spatial autocorrelation shows that the adjacent locations have different values and tend to spread [2]. Characteristics of spatial autocorrelation expressed by Kosfeld, namely:

1. If there is a systematic pattern in the spatial distribution of observed variables, then there is spatial autocorrelation.
2. If the proximity or adjacency between regions closer, it can be said there is positive spatial autocorrelation.
3. Negative spatial autocorrelation illustrates a pattern adjacency unsystematic.
4. The random pattern of spatial data showed no spatial autocorrelation.

Measurement of spatial autocorrelation for spatial data can be calculated using the Moran's Index (Moran), Geary's C, and Tango's excess. In this study, the analysis method is limited only to the

method of Moran's Index (Moran) [2], [10]. This method can be used to detect the onset of spatial randomness. This spatial randomness may indicate clusterization or forming a trend towards space.

D. Spatial Weighted Matrix

Spatial weighted matrix is a matrix that expresses the relationship of the observed region that belongs to $n \times n$ and is denoted by W. The general matrix form of spatial weights (W) shown in (1).

$$w = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{11} & w_{11} & \cdots & w_{11} \\ \vdots & \vdots & \ddots & \vdots \\ w_{11} & w_{11} & \cdots & w_{11} \end{bmatrix} \quad (1)$$

The elements of W above are w_{ij} with i are rows in elements W and j are columns in elements W and are regions around the observation location i. element W above can have two values that are zero and one [11]. Where the value of w_{ij} = 1 for the region adjacent to the location of the observation, while the value w_{ij} = 0 for areas not adjacent to the observation location [12]–[14]. In general there are three types of interaction or border crossing area [15]–[17], namely:

1. Rook Contiguity
   Rook contiguity is the contact of one side with the other side of the neighboring area. The value of each element is that if the location i and j are in contact with the side then w_{ij} = 1. However, if the location i and j are not touching side then w_{ij} = 0. 2
2. Bishop Contiguity
   Bishop contiguity is the juxtaposition of one region with another neighbor. The value of each element is that if the location i and j touch the vertex then w_{ij} = 1. However, if the location of i and j is not touching the vertex then w_{ij} = 0.
3. Queen Contiguity
   Queen contiguity is the contact of the side and corner of the one region with other areas of combined rook contiguity and bishop contiguity. As for the value of each element that is if location i and j touching side or vertex then w_{ij} = 1. However, if location i and j not touching side or corner point then w_{ij} = 0.

E. Moran's Index

The theory of spatial autocorrelation is an important element according to investigation process of geographical spatial from different viewpoints [18]. Moran's I is a development of the Pearson correlation in the univariate data series. Pearson correlation (ρ) between the predictor variables and the response variable with a lot of data n using formula (2).

$$\rho = \frac{\sum_{i,=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i,=1}^{n} (x_i - \bar{x})^2 \sum_{i,=1}^{n} (y_i - \bar{y})^2}} \quad (2)$$

where $\bar{x}$ and $\bar{y}$ the Pearson correlation equation is an average sample of predictor variables and the response. P value is used to measure whether the predictor variables and the response correlated.

The coefficient of Moran's I used to test the spatial dependency or autocorrelation between observations or location [19]–[21]. The test statistic used formula (3) to (10) with the hypothesis for H0 is I = 0 (no autocorrelation between locations), and H1 is I ≠ 0 (autocorrelation between locations).

$$Z_{hitung} = \frac{I - I_o}{\sqrt{\text{var}(I)}} \sim N(0,1) \quad (3)$$

where

$$I = \frac{\mathrm{n}}{\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} w_{\mathrm{ij}}} \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} w_{\mathrm{ij}}(x_{\mathrm{i}} - \bar{x})(x_{\mathrm{j}} - \bar{x})}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_{\mathrm{i}} - \bar{x})^2} \tag{4}$$

$$\mathrm{E}(I) = \mathrm{I}_o = -\frac{1}{\mathrm{n}-1} \tag{5}$$

$$\mathrm{var}(I) = \frac{\mathrm{n}^2 S_1 - \mathrm{n}S_2 + 3S_o^2}{(\mathrm{n}^2 - 1)S_o^2} \tag{6}$$

$$S_1 = \frac{1}{2}\sum_{\mathrm{i}\neq\mathrm{j}}^{\mathrm{n}}(w_{\mathrm{ji}} + w_{\mathrm{ij}})^2 \tag{7}$$

$$S_o = \sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} w_{\mathrm{ij}} \tag{8}$$

$$w_{\mathrm{io}} = \sum_{\mathrm{j}=1}^{\mathrm{n}} w_{\mathrm{ij}} \tag{9}$$

$$w_{\mathrm{oi}} = \sum_{\mathrm{j}=1}^{\mathrm{n}} w_{\mathrm{ji}} \tag{10}$$

Where xi are variable data to location-i (i = 1, 2, ..., n), *xj* as the variable data to location-j(j = 1, 2, ..., n), $\bar{x}$ for Data Average, var (I) ar varians Moran's I, and E(I) is an expected value Moran's I. Decision-making reject Ho if $\left|Z_{hitung}\right| > Z_{\alpha/2}$. The value of the index I is between -1 and 1. If I> Io, the data has a positive autocorrelation, if I < Io, the data has negative autocorrelation, and Moran index value is zero indicating no groups. Moran index value does not guarantee the accuracy of measurement if the weighting matrix used are not standardized weighting.

F. Spatial Autoregressive Model (SAR)

Spatial Model Autoregressive is a model that combines a simple regression model with spatial lag dependent variable using cross-section data [3], [10]. Autoregressive spatial models were formed when W2 = 0 and λ = 0, so that this model assumes that the autoregressive process only on the response variable [10]. The general model SAR is shown in (11).

$$y = \rho W_1 y + X\beta + \varepsilon \tag{11}$$

where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$, $\boldsymbol{y}$ is bounded variable vector size $n \times 1$, ρ are spatial autocorrelation coefficient on dependent variable, $\boldsymbol{W}$ is a spatial weighted matrix of $n \times n$ size, $\boldsymbol{X}$ for free variable matrix size $n \times (k + 1)$, β represent a vector of regression coefficient parameters measuring $k \times 1$, and ε is a vector error free autocorrelation size $n \times 1$. While estimation of β parameter in Spatial Autoregressive Model obtained by using likelihood maximum method is as in (12).

$$\beta = (X'X)^{-1}X'(y - \rho W y) \tag{12}$$

This model is the development of the first order autoregressive model, where the response variable in addition affected by the lag response variable itself is also influenced by the predictor variables. Autoregressive process also has similarities with the analysis of the time series as the first order autoregressive spatial models.

III. Methods

The analytical method used is the method of spatial regression analysis, namely Spatial Autoregressive Model (SAR). The method of research is done using algorithms (Fig.1).

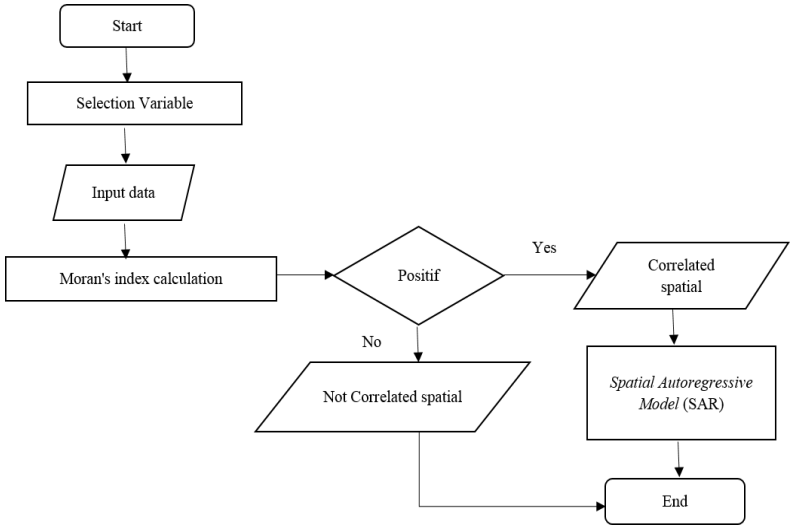


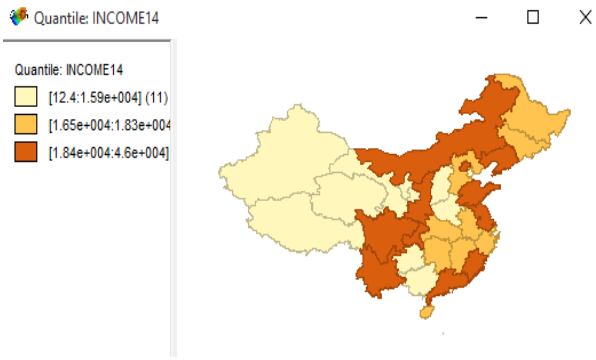
Fig. 1. Flowchart of Research Method

The data used is secondary data, the existing data on the website of the National Bureau of Statistics of China in 2014. The variables used are Per Capita Disposable Income Nationwide (yuan), Foreign Exchange Earnings from International Tourism (USD million), Total Value of Exports of operating units (1,000 US dollars), Registered Unemployed Persons in Urban Area (10000 persons), Number of Overseas Visitor Arrivals (million person-times), Number of Industrial Enterprises above Designated Size (unit).

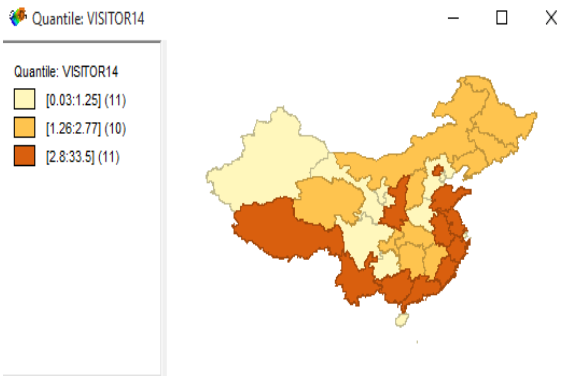
IV. Result and Discussion

A. Data Exploration using Thematic Map

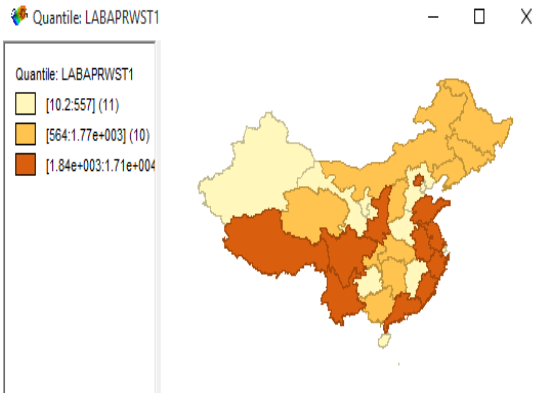
Before analyzing the Moran's Index, The research aim to explore the thematic map of the variables. This is to see whether there was a spatial pattern or no. Fig 2 are the results of the thematic map on each variable. It shows that there was a spatial correlation pattern between regions. It is indicated that was a positive spatial correlation between regions because the same colors are closed each other this notify that same value are closed each other. These diagnostics are not certainly true before the research continued to calculate the Moran's Index between the variables. The next paragraph explains this hypothesis.



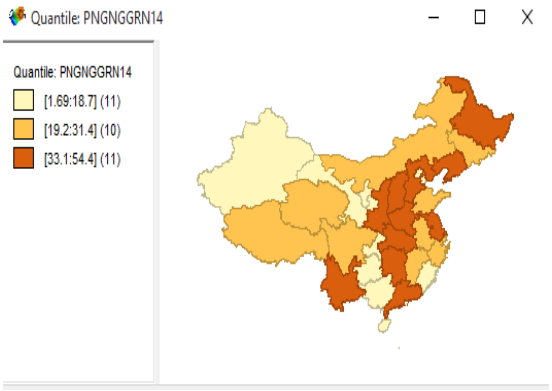
(a) *Mapping Per Capita Disposable Income Nationwide (yuan)*



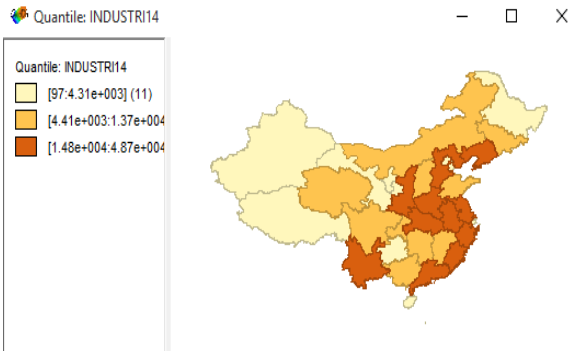
(b) *Mapping Number of Overseas Visitor Arrivals(million person-times)*



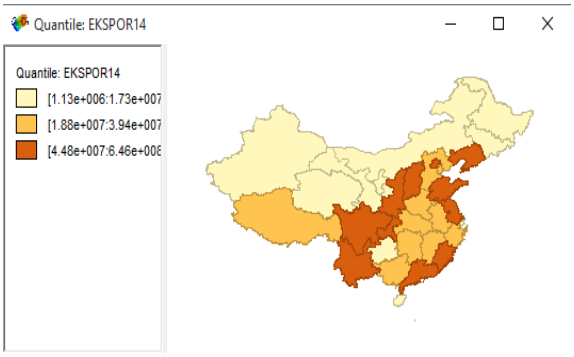
(c) *Mapping Foreign Exchange Earnings from International Tourism (USD million)*



(d) *Mapping Registered Unemployed Persons in Urban Area (10000 persons)*



(e) *Mapping Number of Industrial Enterprises above Designated Size (unit)*

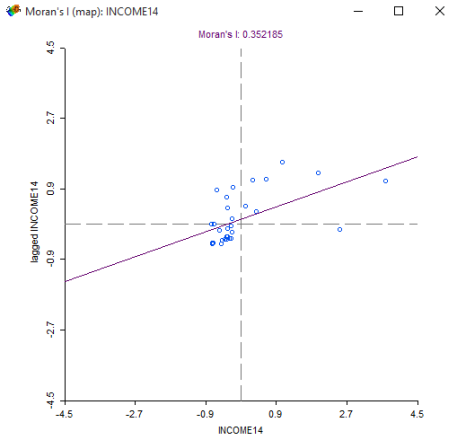


(f) *Mapping Total Value of Exports of operating units (1,000 US dollars)*

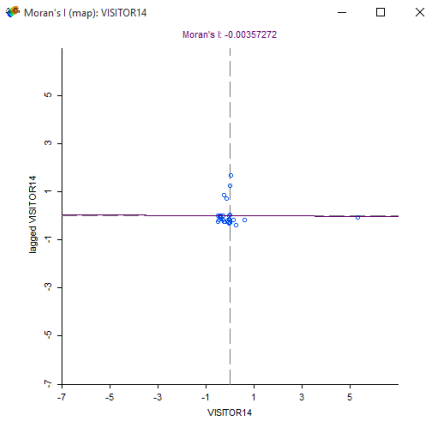
Fig. 2.Thematic Map

B. Moran's Index

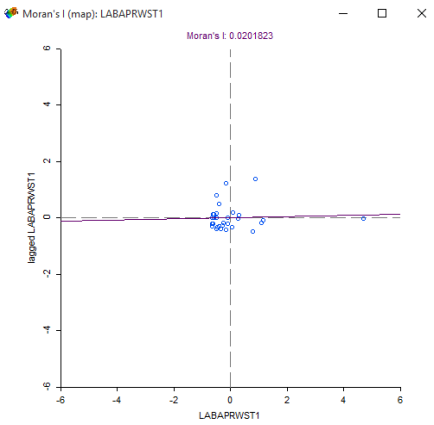
Furthermore, for a measure of certainty whether there is a correlation between the value of observations relating to spatial locations on the same variables that indicate the similarity value of the locations are adjacent and tend to cluster can be done through the measurement of spatial autocorrelation in this case used Moran's Index method. The results are shown in Fig.3.



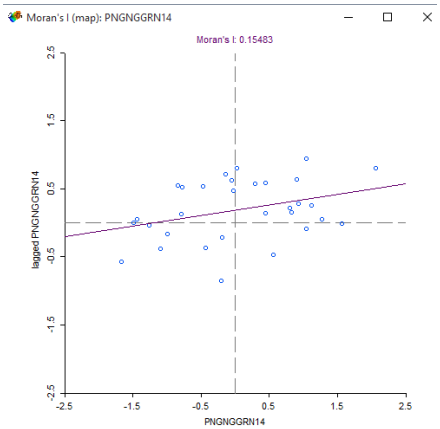
(a) *Moran's Index Per Capita Disposable Income Nationwide (yuan)*



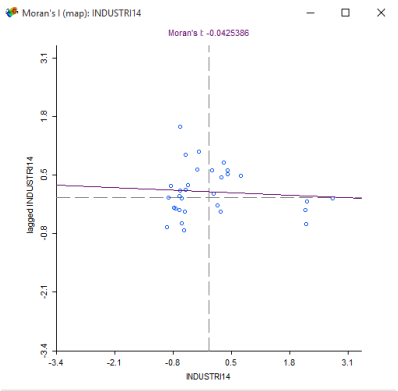
(b) *Moran's Index Number of Overseas Visitor Arrivals (million person-times)*



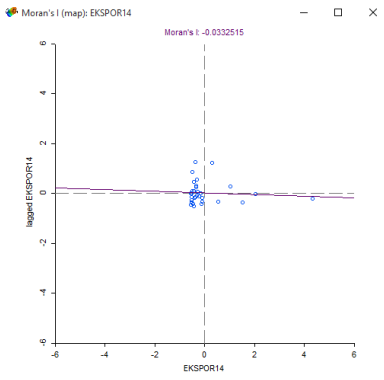
(c) *Moran's Index Foreign Exchange Earnings from International Tourism*



(d) *Moran's Index Registered Unemployed Persons in Urban Area (10000 persons)*



(e) *Moran's Index Number of Industrial Enterprises Above Designated Size (unit)*



(f) *Moran's Index Total Value of Exports of operating units (1,000 US dollars)*

Fig. 3. Moran's Index

From the calculation using Geoda, found that there was three variables that have a positive Moran's Index value, namely: Nationwide Per Capita Disposable Income (yuan) with the value I = 0.352, Foreign Exchange Earnings from International Tourism to the value I = 0.02, and Registered Unemployed persons in Urban Area (10000 persons) with a value of I = 0,154. Moran positive index value indicates that the variable has an element of proximity or adjacency between regions closer.

C. SAR Model

After the Moran index value calculation, the next analysis was using SAR Model. Table 1 shows the output result of SAR model which has indicated that the variable export was not significant. After spending the variables are not significant, then get the output with all the significant variables are as shown in Table 2.

Table 1. Output SAR with all variables

Variable	Coefficient	Std. Error	Z-Value	Probability
W_Income	-0.18	0.149	-1.211	0.225
Constant	14882.95	3005.09	4.952	0.000
Industry	-0.415	0.1764	-2.353	0.018
Tourism	5.827	1.053	5.590	**0.000**
Export	4.527	2.751	1.645	0.099
UnEmployment	269.41	96.986	2.777	0.0054
Visitor	-3167.35	521.096	-6.078	0.000

Table 2. Output SAR with significant variables.

Variable	Coefficient	Std. Error	Z-Value	Probability
W_Income	-0.09913	0.1559	-0.635	0.524
Constant	16881.27	2995.897	5.634	0.000
Tourism	5.608	0.9178	6.111	0.000
Visitor	2448.09	505.1265	-4.846	0.000

Based on the Table 1 and 2 can be obtained Model Spatial Autoregressive (SAR) for Nationwide Per Capita Disposable Income (yuan) China in 2014 was as shown in (12).

$$y_i = -0.09913 \sum_{j \neq i}^{N} w_{ij} y_j + 5{,}6087\ X_1 - 2448{,}09\ X_2 + \varepsilon_i \quad (12)$$

At level α = 5%, it can be said that the Nationwide Per Capita Disposable Income (yuan) in a regions i affected by the number of Per Capita Disposable Income Nationwide (yuan) of their neighbors. Interpretation of Spatial Autoregressive models are for each additional 1% Foreign Exchange Earnings from International Tourism alone will increase the average Per Capita Disposable Income Nationwide (yuan) by 100 (e5,6087-1)%. Variable Number of Overseas Visitor Arrivals (million person-times) had a regression coefficient is negative then for every increase of 1% Number of Overseas Visitor Arrivals (million person-times) will lower the Nationwide Per Capita Disposable Income (yuan) by 100 (1 – e-2448,09(1)) %.

The coefficient of determination for the model Spatial Autoregressive (SAR) of 0.5865 means that the model can explain the diversity Nationwide Per Capita Disposable Income (yuan) amounted to 58.65%, while the rest is explained by variables outside the model.

V. Conclusion

Data Per Capita Disposable Income Nationwide (yuan) in China from year 2014 can be modeled using SAR and there was two factors used significantly influence Nationwide Per Capita Disposable Income (yuan). It is Foreign Exchange Earnings from International Tourism and Number of Overseas Visitor Arrivals (million person-times). In addition, this study can also be used as the basis for the government in China in order to enhance the Foreign Exchange Earnings from International Tourism

in order to increase the average value of Per Capita Disposable Income Nationwide (yuan) and pressing factor Number of Overseas Visitor Arrivals (million person-times).

Acknowledgements

We thank to Universitas Islam Indonesia which has funded this research. We are also grateful to all parties for support this research, providing data and making this research easier.

References

[1] NBS, "National Bureau of Statistics of China," *National Bureau of Statistics of China*, 2017.

[2] A. Fotheringham and P. Rogerson, *The SAGE Handbook of Spatial Analysis*. 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications, Ltd, 2009.

[3] L. Anselin, *Spatial Econometrics: Methods and Models*, vol. 4. Dordrecht: Springer Netherlands, 1988.

[4] Z. Yu *et al.*, "Geometric Algebra Model for Geometry-oriented Topological Relation Computation," *Trans. GIS*, vol. 20, no. 2, pp. 259–279, 2016.

[5] Z. Mustafa, J. Flores, J. M. Cotos, and E. Abad, "New Developments in the use of Spatial Technology in Archaeology, Sample case: Rocha Castle System," *Int. J. Adv. Stud. Comput. Sci. Eng.*, vol. 5, no. 11, p. 186, 2016.

[6] K. Thomas *et al.*, "A simple approach for a spatial terrestrial exposure assessment of the insecticide fenoxycarb, based on a high-resolution landscape analysis," *Pest Manag. Sci.*, vol. 72, no. 11, pp. 2099–2109, 2016.

[7] R. Zuo, E. J. M. Carranza, and J. Wang, "Spatial analysis and visualization of exploration geochemical data," *Earth-Science Rev.*, vol. 158, pp. 9–18, 2016.

[8] L. G. Roser, L. I. Ferreyra, B. O. Saidman, and J. C. Vilardi, "EcoGenetics: an R package for the management and exploratory analysis of spatial data in landscape genetics," *Mol. Ecol. Resour.*, 2017.

[9] W. Yu, "Spatial co-location pattern mining for location-based services in road networks," *Expert Syst. Appl.*, vol. 46, pp. 324–335, 2016.

[10] M. D. Ward and K. S. Gleditsch, *An Introduction to Spatial Regression Models in the Social Sciences*. Los Angeles: SAGE Publications, Ltd, 2008.

[11] J. P. LeSage, "The theory and practice of spatial econometrics," *Univ. Toledo. Toledo, Ohio*, vol. 28, p. 33, 1999.

[12] M. Deng, Q. Liu, J. Wang, and Y. Shi, "A general method of spatio-temporal clustering analysis," *Sci. China Inf. Sci.*, vol. 56, no. 10, pp. 1–14, 2013.

[13] C. Morrison, W. R. Ponicki, P. J. Gruenewald, D. J. Wiebe, and K. Smith, "Spatial relationships between alcohol-related road crashes and retail alcohol availability," *Drug Alcohol Depend.*, vol. 162, pp. 241–244, 2016.

[14] J. W. Lichstein, T. R. Simons, S. A. Shriner, and K. E. Franzreb, "Spatial autocorrelation and autoregressive models in ecology," *Ecol. Monogr.*, vol. 72, no. 3, pp. 445–463, 2002.

[15] X. Ye, "Spatial econometrics," *Int. Encycl. Geogr.*, 2016.

[16] G. K. A. Harvey, T. A. Nelson, C. H. Fox, and P. C. Paquet, "Quantifying marine mammal hotspots in British Columbia, Canada," *Ecosphere*, vol. 8, no. 7, 2017.

[17] A. Prahutama and A. Hoyyi, "Modeling of Malaria Spread in Central Java Using Spatial Regression," *Adv. Sci. Lett.*, vol. 23, no. 7, pp. 6537–6540, 2017.

[18] Y. Chen, "New approaches for calculating Moran's index of spatial autocorrelation," *PLoS One*, vol. 8, no. 7, p. e68336, 2013.

[19] R. P. Haining, *Spatial data analysis: theory and practice*. Cambridge, UK ; New York: Cambridge University Press, 2003.

[20] D. M. Lambert, J. P. Brown, and R. J. G. M. Florax, "A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application," *Reg. Sci. Urban Econ.*, vol. 40, no. 4, pp. 241–252, 2010.

[21] Y. M. Zhukov, "Applied Spatial Statistics in R, Section 2 Spatial Autocorrelation," 2010.